Data Vault – Ein Leben zwischen 3NF und Star

Michael Klose
Logica Deutschland GmbH & Co. KG, now Part of CGI
Sulzbach (Taunus)

Schlüsselworte
Data Vault, Datenmodellierung, Enterprise Datawarehouse, Compliance

Einleitung
Die klassische Modellierung des Core-Datawarehouse mit dritter Normalform beweist sich in der Praxis bei steigendem Datenvolumen und schnellen Go-to-Market Anforderungen als nicht mehr zeitgemäß.
Data Vault stellt eine Modellierungsvariante dar, welche sich zwischen 3 NF und Star Schema Modellierung positioniert. Wenige Tabellentypen (Hub, Sat, Link) und eine fachlich orientierte Abbildung des Modells vereinfachen die Modellierung.
Im Vortrag wird anhand von Praxisbeispielen aus Projekten gezeigt welche verschiedenen Modellierungsvarianten für gleiche Anforderungen gewählt werden können. Weiterhin wird dargestellt wie Compliance Anforderungen im Zusammenspiel von Raw und Business Data Vault erfüllt werden können.

Was ist Data Vault?

Dan Linstedt der Urheber von DataVault sagt:

„DataVault ist eine detailorientierte, historisch genaue und eindeutig nachvollziehbare Anordnung von normalisierten Tabellen.“

Dieser sieht DataVault als logische Fortsetzung der Entwicklung im Data Warehousing. DataVault ist ein Hybrid-Ansatz aus 3NF und Star-Schema und bietet neben einer impliziten Historisierung und Versionierung eine hohe Flexibilität bei Änderungen im Datenmodell.

DataVault betrachtet die Modellierung aus der Sicht von Geschäftsdaten und beginnt damit Geschäftsobjekte wie Mitarbeiter, Abteilungen, Produkte, und Bestellungen zu identifizieren. Diese werden in sogenannten Hubs gespeichert. Zwischen Hubs werden die Beziehungen ausschließlich durch Link-Tabellen realisiert. Alle Informationen zu den Geschäftsobjekten und den Verbindungen zwischen ihnen werden in Satelliten-Tabellen gespeichert.

Datensätze werden laut der reinen DataVault Theorie weder geändert, noch gelöscht. Geänderte Daten werden durch einen neuen Datensatz mit einem aktuellen Load_Date eingepflegt. Gültigkeits-zeiträume werden häufig mittels Valid_From und Valid_To gespeichert. Ein Valid_From ist dann sinnvoll, wenn Daten bereits gültig sein können, bevor sie ins DWH geladen werden. Dies kann bspw. bei monatlichen Uploads der Fall sein. Das Valid_To Datum ist für eine Markierung von ungültigen Daten notwendig, ohne dass diese aus der Historie entfernt werden.

Business Data Vault

Man kann zwischen Raw DataVault und Business DataVault unterscheiden. Bei Raw DataVault werden alle Daten aus den Quellsystemen aufgenommen und auditierbar sowie historisiert verwaltet. Die Anwendung von Business Rules wird nachgelagert. Business DataVault hingegen enthält Daten, auf die bereits Business Rules (Geschäftsregeln) angewandt wurden und bereinigt sind.

Daten im Business DataVault stammen dabei immer aus einem Raw / Operational DataVault. Dabei existieren zwei grundsätzliche Ansätze für das Design: Zum einen kann man zwei Schemata anlegen, eines für das Raw DataVault Modell und eines für das Business DataVault Modell. Alternativ kann das Raw DataVault um Business Hubs, Links und Satelliten erweitert werden.

So können bspw. im Raw DataVault Kundenadressen aus zwei unterschiedlichen Quellsystemen abgelegt werden. Anschließend kann durch einen Master Data Management-Algorithmus (MDM), die aktuell gültige Adresse bestimmt und in das Business DataVault Schema übernommen bzw. in einen Business Satelliten abgelegt werden.

Herausforderungen 3NF

Im Folgenden wird DataVault anhand eines Beispiels veranschaulicht.

Ein Unternehmen hat Mitarbeiter, die Abteilungen zugeordnet werden. Zu den Mitarbeitern liegen die Informationen Personalnummer, Name und Vorname vor. Abteilungen haben einen Namen und eine Abteilungsnummer. Die Modellierung des Sachverhalts mittels 3NF gestaltet sich wie folgt:

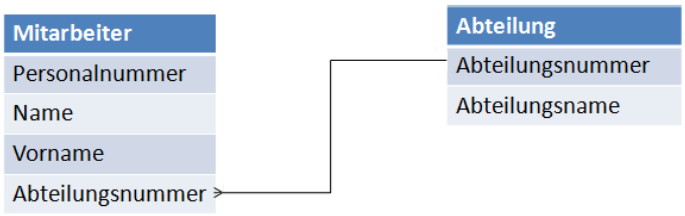

Abb. 1: Zuordnung von Mitarbeitern zu Abteilungen (Beispiel, 3NF)

Beide Tabellen haben jeweils einen Primärschlüssel in der Form der Personalnummer bzw. der Abteilungsnummer. Die Beziehung zwischen den Tabellen wird mithilfe eines Fremdschlüssels über die Abteilungsnummer realisiert.

Typische Herausforderungen, die bei der Modellierung mit 3NF entstehen können:

Was passiert wenn der Abteilung auch ein Abteilungsleiter zugeordnet werden soll?
Es muss eine neue Spalte in die Tabelle Abteilung eingefügt werden.
Diese braucht einen Fremdschlüssel zur Personalnummer in der Tabelle Personal

Was passiert wenn ein Mitarbeiter mehreren Abteilungen zugeordnet werden kann?
Zwischen Mitarbeiter und Abteilung muss eine Zuordnungstabelle erzeugt werden, die zwei Fremdschlüssel benötigt.
Die Spalte Abteilungsnummer in der Tabelle Mitarbeiter wird damit obsolet.

Was passiert wenn dem Mitarbeiter sein Vorgesetzter zugeordnet werden soll oder sogar mehrere Vorgesetzte haben kann?
Im ersten Fall (ein Vorgesetzter) wird die neue Spalte Vorgesetzter in die Tabelle Mitarbeiter eingefügt.
Diese bekommt einen Fremdschlüssel, der wiederum auf die Tabelle Mitarbeiter verweist.
Im zweiten Fall (mehrere Vorgesetzte) wird, wie bei der Zuordnung von Mitarbeitern, zu mehreren Abteilungen eine Zuordnungstabelle benötigt, die zwei Fremdschlüssel zur Personalnummer hat.
Wenn die Fälle nacheinander auftreten wird beim Modellieren des zweiten Falls die Spalte Vorgesetzter obsolet.

Was passiert, wenn nachvollziehbar sein soll, in welchem Zeitraum der Mitarbeiter in welcher Abteilung war?
Es wird eine Tabelle Arbeitshistorie eingefügt, die auf Mitarbeiter sowie Abteilung referenziert.
Zusätzlich enthält die Tabelle den Arbeitszeitraum und benötigt einen künstlichen Schlüssel.

Modellierung Data Vault

Das angeführte Beispiel wird mithilfe von DataVault wie folgt modelliert:

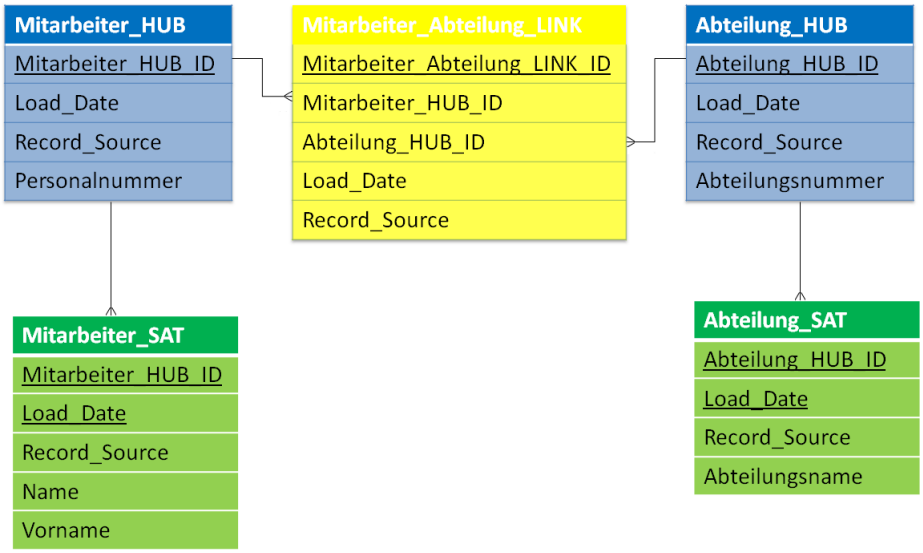

Abb. 1: Zuordnung von Mitarbeitern zu Abteilungen (Beispiel, DataVault)

Die in verwendeten Konstrukte und deren Beziehungen untereinander werden im Folgenden erläutert.
.

HUB
Ein Hub enthält einen eindeutigen Business Key, der zu einem Typ von Geschäftsobjekten gehört. Im Beispiel sind das zum einen Personalnummern und zum anderen Abteilungsnummern. Diese Business Keys werden mit einem Unique-Schlüssel versehen. Ein Hub enthält zusätzlich immer einen Primärschlüssel, der in der Regel ein künstlich erzeugter technischer Schlüssel ist. Im Load_Date wird eingetragen, wann der Business Key erstmals geladen wurde. In die Record_Source wird der Ursprungsort des Datensatzes abgelegt. Durch Load_Date und Record_Source werden die Daten nachvollziehbar und auditierbar.

LINK
Ein Link stellt Beziehungen zwischen Hubs her. Dieser enthält neben seinem künstlich erzeugten Primärschlüssel die Primärschlüssel der verbundenen Hubs als Fremdschlüssel. Zudem wird ein Unique-Schlüssel über die Fremdschlüsselspalten erstellt. Hubs enthalten ebenso wie Links aus Gründen der Nachvollziehbarkeit das Load_Date und die Record_Source. Mithilfe des Links können Personalnummern und Abteilungsnummern einander zugeordnet werden. Um die in der Abbildung noch fehlenden Informationen zu den Mitarbeitern und Abteilungen darzustellen, wird das dritte Grundelement von DataVault, der Satellit, benötigt.

SATELLIT
Ein Satellit wird immer genau einem Hub oder einem Link zugeordnet. Der Primärschlüssel setzt sich aus der Kombination des Schlüssel des Hubs oder Links zu dem der Satellit gehört sowie dem Load_Date zusammen. Zusätzlich enthält der Satellit noch die Record Source und die Detailinformationen (Attribute).Mithilfe von Satelliten können bspw. die Namen zu den Mitarbeitern und Abteilungen in der Datenbank verwaltet werden. Jedem Hub oder Link können mehrere Satelliten angehängt werden.

Gültigkeit von Datensätzen

DataVault verfolgt eine No-Update-Strategie. Aus diesem Grund ist das Load_Date in den Satelliten Bestandteil des Primärschlüssels. Ändert sich z. B. ein Abteilungsname, wird in den Satelliten eine neue Zeile eingetragen, die den neuen Namen enthält. Dadurch kommt die Abteilungsnummer zwar doppelt im Satelliten vor, jedoch mit unterschiedlichem Load_Date. Der aktuell gültige Datensatz wird immer am aktuellsten Load_Date erkannt. Es werden nur reine Inserts durchgeführt und kein Merge realisiert, so dass sich die Ladeperformance verbessert. Hierdurch wird eine implizite Versionierung / Historisierung erreicht, da alle Änderungen nachvollzogen werden können.

Kontaktadresse:
Michael Klose
Logica Deutschland GmbH & Co. KG | Now part of CGI
Am Limespark 2
D-65843 Sulzbach (Taunus)

Telefon: +49 (0) 171-977 90 99
E-Mail Michael.Klose@logica.com
Internet: www.logica.com | www.cgi.com